A simple transformation independent method for outlier definition

Johansen, Martin Berg; Christensen, Peter Astrup

Published in:
Clinical Chemistry and Laboratory Medicine

DOI (link to publication from Publisher):
10.1515/cclm-2018-0025

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Martin Berg Johansen and Peter Astrup Christensen*

A simple transformation independent method for outlier definition

https://doi.org/10.1515/cclm-2018-0025
Received January 9, 2018; accepted February 22, 2018; previously published online April 10, 2018

Abstract

Background: Definition and elimination of outliers is a key element for medical laboratories establishing or verifying reference intervals (RIs). Especially as inclusion of just a few outlying observations may seriously affect the determination of the reference limits. Many methods have been developed for definition of outliers. Several of these methods are developed for the normal distribution and often data require transformation before outlier elimination.
Methods: We have developed a non-parametric transformation independent outlier definition. The new method relies on drawing reproducible histograms. This is done by using defined bin sizes above and below the median. The method is compared to the method recommended by CLSI/IFCC, which uses Box-Cox transformation (BCT) and Tukey's fences for outlier definition. The comparison is done on eight simulated distributions and an indirect clinical datasets.
Results: The comparison on simulated distributions shows that without outliers added the recommended method in general defines fewer outliers. However, when outliers are added on one side the proposed method often produces better results. With outliers on both sides the methods are equally good. Furthermore, it is found that the presence of outliers affects the BCT, and subsequently affects the determined limits of current recommended methods. This is especially seen in skewed distributions. The proposed outlier definition reproduced current RI limits on clinical data containing outliers.
Conclusions: We find our simple transformation independent outlier detection method as good as or better than the currently recommended methods.

***Corresponding author: Peter Astrup Christensen,** Department of Clinical Biochemistry, Aalborg University Hospital, Hobrovej 18-22, 9000 Aalborg, Denmark, Phone: +45 97649000,
E-mail: Peter.Christensen@rn.dk
Martin Berg Johansen: Unit of Clinical Biostatistics, Aalborg University Hospital, Aalborg, Denmark

Keywords: binning; non-parametric; outlier; reference interval; transformation independent; Tukey's fences.

Introduction

Outliers and reference intervals

Medical laboratories establishing or verifying reference intervals (RI) are facing the challenge of detecting erroneous values in the datasets. These outlying observations can be caused by experimental error or measurement variability. If a possible experimental error affecting an observation cannot be identified, medical laboratories often rely on a variation of statistical methods to identify these outliers. Definition and elimination of outliers is crucial in all sorts of distributions in clinical biochemistry, as the inclusion of outliers affects the determination of limits. Most prominent this is seen in skewed distributions where inclusion of just a few outlying observations can have huge impact on the determination of the reference limits. An ideal algorithm for outlier detection should identify any number of observations distant from other observations regardless of probability distribution [1]. A diverse range of methods has been developed for identifying outlying observations. In contrast to an ideal algorithm, these methods often rely on assumptions about the underlying distributions.

Present methods

Several methods assumes that the data originates from a normal distribution. In clinical biochemistry, these methods include Dixon's Q test and Grubbs's test [2, 3], which are both essentially designed to detect only one outlier. Other used methods are elimination of extremes; e.g. values which fall outside mean ± 3 or 4 standard deviations (SD) [4–8].

Tukey's fences are also a very popular way of defining outliers [9]. It relies on the central part of the distribution by using the interquartile range (IQR) defined as the third quartile minus the first quartile. Based on this the outlier limits are defined by Lower limit = first quartile − 1.5 ∗ IQR and Upper limit = third quartile + 1.5 ∗ IQR. This procedure is reproducible but highly sensitive to skewness, and the

method is only recommended for symmetrical distributions. Furthermore, the outcome of outlier removal using Tukey's fences is dependent on sample size, as also the estimation of the population quantiles are dependent on sample size [10, 11]. These methods can also be used on skewed distributions if the non-normal data are transformed to normality before use of the test. However, for applications in clinical biochemistry Dixon's Q test is considered reasonably insensitive to distribution type [12].

To facilitate use of Tukey's fences in skewed distributions it is used in conjunction with several different transformations, examples being the natural logarithmic transformation or the Box-Cox transformation (BCT) (equation 1) [13].

$$T(Y)=\begin{cases}\dfrac{Y^{\lambda}-1}{\lambda} & \lambda\neq 0\\ \ln(Y+c) & \lambda=0\end{cases} \quad (1)$$

Here, λ is a parameter that determines the shape of the transformation and c is a location constant. λ is estimated using the maximum likelihood estimation in the original data [13].

Finally, one can use the more advanced methods like BCT followed by adjustment for remaining kurtosis [14]. After transformation, Tukey's fences are applied to define outliers. Or the choice can be the yet more statistically complex robust estimator [15]. Despite the availability of advanced calculus programs, these procedures require knowledge and care in order to be used sensibly.

In practice, a histogram is often used to perform a visual assessment for outliers. Clearly, this method does not depend on assumptions regarding the distribution but it is not reproducible, as it depends on the calculus program used to draw the histogram. Selecting the bin size and anchor point for the histogram can have large effects on the visual outcome. However, it is probably still a widely used method amongst doctors in clinical biochemistry in the everyday work in the laboratory. Several different proposals for determining the optimal bin size have been developed. They all rely on assumptions on the underlying distribution, and most of them work best for normal distributions.

The Freedman-Diaconis rule suggests that using a bin size h defined as

$$h=\frac{2*\mathrm{IQR}}{\sqrt[3]{\mathrm{n}}} \quad (2).$$

This bin size minimizes the difference between the areas of the empirical and the theoretical probability distribution functions. It only requires the mild assumption that the underlying distribution is not uniform. Furthermore, it is not very sensitive to outliers in the data, as it does not depend on the standard deviation or total range. Drawing a histogram with this bin size gives a rough sense of the density of the underlying distribution of the data and estimates the probability density function [16].

Current recommendations

The current recommendations (EP28-A3C) published by the Clinical and Laboratory Standards Institute (CLSI) and the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) mention the following procedures when determining RIs [17]:

First, it is recommended to do a visual examination of the frequency distribution. Secondly, one of the following two procedures could be used for identifying outlying observations:

1. Dixon's Q test. Using this procedure on the least extreme outlier as if it was the only outlier and subsequently also remove more extreme outliers, allows for defining more than one outlier on the same side of the distribution, or
2. Box-Cox transformation in conjunction with Tukey's fences (BCT). The BCT removes skewness. It is recommended to repeat this method until no further outliers are detected.

Drawbacks for recommended methods

Dixon's Q test is very simple to use, but it can fail to detect outliers when these are masked by spacing [18]. Especially, it can have problems when the distribution is not normal, as it is an underlying assumption. For the second method, the presence of outliers in the dataset has influence on the estimated λ-parameter, and thus influence the transformation.

Description of a proposed method

With very large sample sizes as is common in, e.g. engineering sciences, outliers can be defined as data not connected to the main probability distribution [19]. However, for applications like RIs in clinical biochemistry where limited number of data are available this method is not directly transferable. Additionally before evaluating the histogram, the skewness of clinical biochemistry data should be taken into account. This can be done in a histogram by using two modified Freedman-Diaconis bin sizes, one on each side of the median.

$$h_1 = \frac{4*(\text{median} - 1\text{st quartile})}{\sqrt[3]{n}} \quad (3)$$

$$h_2 = \frac{4*(3\text{rd quartile} - \text{median})}{\sqrt[3]{n}} \quad (4).$$

This ensures a greater precision of the estimated probability density function in each tail of the distribution, which is important for the outlier definition purpose. It also renders the median as a natural anchoring point for the histogram. Plotting the histogram with bin size h_1 below the median and h_2 above the median allows for a new method for defining outliers. Extreme data larger or smaller than points where the probability density function becomes zero are not connected to the main part of the distribution, and could therefore be considered outliers.

As we find our outlier definition with this modified Freedman-Diaconis binning (FDB) method intuitive and simple we compared it to the BCT method which is the current recommendation of CLSI/IFCC [17]. We compared the outlier definition in simulations on eight different distributions with different methods of adding outliers and different sample sizes. Subsequently, we evaluated the methods on clinical data by using an indirect approach for evaluating RIs.

Materials and methods

Generation of random data

Calculations, statistical evaluations, graphical representations and data generation were made in Rstudio (Version 1.0.136 with R Version 3.3.2). The following packages were installed (plyr, dplyr, ggplot2, forecast, gsubfn, LaplacesDemon). Rstudio with R and the installed R packages were used to simulate the following distributions:

Normal (mean = 10, SD = 1); half-normal; log-normal; χ^2 (df [degrees of freedom] = 1, 4 and 8); and χ^2 (df = 1, ncp [non-centrality parameter] = 10). Square root normal distributions were generated by squaring the simulated normal distributions (before and after addition of outliers to the normal distributions). For each distribution type we generated 1000 replicates of samples of sizes n = 30, 60 and 120. All simulated values were rounded to the fourth decimal. To ensure strictly positive numbers in the distributions, the values rounded to 0 are replaced with 10^{-4}.

Adding outliers on one side

Outliers were added using a method similar to previous published methods [1]. Each sample had one through six random values replace by outliers. Outliers were added as a uniform random variable on the upper side of the distribution in the probability range (0.35–0.005%) corresponding to mean + 2.7–3.9 SDs in the normal distribution. This gives on average <1 outlier in samples without outliers added for sample size n = 120.

Adding outliers on both sides

For distribution types: Normal; square root normal; χ^2 (df = 8); χ^2 (df = 1, ncp = 10) outliers were also added on both sides. The method was similar to the one-sided model, but c outliers were added in the same probability range with a random number of outliers p to the lower side and $c–p$ added to the upper side of the distribution.

Clinical data: Patient results were extracted from the laboratory information system (LIS) using methods similar to our recently published method [20]. Included test results were all analyzed at Aalborg University Hospital in the period of (21.08.2017–24.09.2017). The following selection criteria were used: Each unique patient had only one request for biochemical testing in the regional LIS within 1.5 years retrospective of an included sample and the retrieved data from the LIS are all from outpatients consulting general practitioners. In total just under 23,000 patient results were extracted. The results are distributed with approx. 54–2125 per stratification. Included tests were (alanine transaminase, albumin, pancreatic amylase, alkaline phosphatase, bilirubin, calcium, creatine kinase, creatinine, iron, γ-glutamyl-transferase, potassium, sodium and urate).

Statistical evaluation

Determination of RIs were done with the non-parametric method using the 2.5th and 97.5th percentiles for sample sizes n = 60 and 120, and using the 5th and 95th percentiles for sample size n = 30. If the rank values are not integers, interpolation was done between rank values on each side [17]. Root mean squared error (RMSE) is calculated to the non-parametric limits of the distribution without outlier replacement.

For outlier elimination with the BCT method, the sample data was transformed using the BCT. Subsequently, Tukey's fences were applied to define outliers. The sample data without outliers were then back transformed to the original scale and reference values were determined with the described non-parametric method.

For the FDB method, the bin sizes were calculated according to formulas (3) and (4) and rounded to five decimals. Frequency distributions were the built from the median using h_1 bin size below the median and h_2 above the median. Starting from the median the lower outlier limit is defined as the first empty bin (frequency = 0) below the median. Similarly, the upper outlier limit is defined as the first empty bin above the median. Sample data outside the limits are defined as outliers and removed. This renders the sample data as one connected frequency distribution given the defined bin sizes.

Ethics

The study was a technical and quality investigation in accordance with the guidelines of the Northern Denmark Regional Science and Ethics Committee.

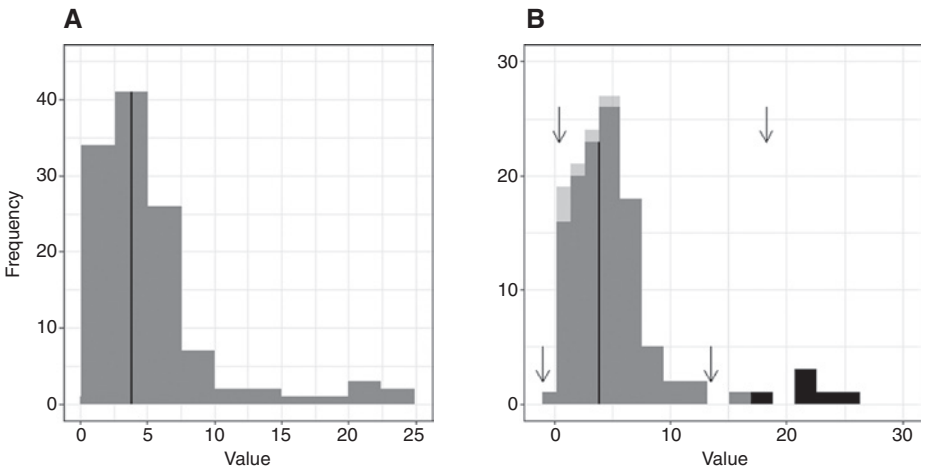

Figure 1: Histograms of a sample from χ^2 (df = 4) distribution with 5% outliers.
Each histogram outlines the same sample (n = 120). The black line indicates the median. (A) Default histogram as Microsoft Excel® would produce it; (B) Histogram drawn using one FD bin size for each half of the diagram. Light gray boxes indicate locations of values, which were replaced by outliers (black boxes). The arrows in the top of the histogram indicates outlier fences as defined with the BCT method. The arrows in the bottom of the histogram indicate the first empty bin larger and smaller than the median. Defining the FDB outlier limits.

Results

Results from simulation experiments shown here are all from (n = 120). Results from (n = 30 and 60) are shown in the Supplementary Data. Figure 1A shows a histogram of one simulated sample from the χ^2 (df = 4) distribution containing 5% outliers. The plot is rendered using the default settings, as it would be in Microsoft Excel® (number of bins defined by $\sqrt{n}$). The histogram has the first bin from 0 to 0.1029 and thereafter a bin size of 2.48, which in this case gives a complete connected frequency distribution. Calculating the bin sizes as defined in equation (3) and (4) gives a bin size of 1.21 below the median and 1.88 above the median. Plotting the histogram using these bin sizes is shown in Figure 1B. The bin locations of the six random values which were replace by outliers are shown in light gray. The added outliers are located in the black bins. In this sample, the positions of the outlier fences with the BCT method are shown as arrows in the top of the histogram. Near the lower limit, the BCT method excludes one non-outlier, and near the upper limit it includes one of the added outliers. Similarly, the positions of the FDB outlier fences are shown as arrows in the bottom of the histogram. This method removes all six outliers and one non-outlier near the upper limit and none near the lower limit.

Simulation experiments

Violin plots are used to show summaries of the simulation experiments. The violin plots show the probability density function of the non-parametric calculated RI limits. The average percentage of detected outliers is indicated in the plots.

Comparing the methods for outlier definition, the BCT method generally defines fewer outliers when none are added. When outliers are added on one side, the FDB method often results in limits closer to the theoretical limits. Especially, on the side where no outliers are added. Here the FDB method renders the limits unaffected whereas the BCT outlier definition affects the limit (Table 1 and Supplementary Tables 1 and 2). Figure 2 shows violin plots of the normal (A) and the χ^2 (df = 1, ncp = 10) (B) distribution, respectively. For these two distributions, the FDB method performs better without outliers added. The distributions of the calculated upper limits are very similar. In contrast to this, the lower limit has shifted when the BCT method is used whereas the FDB method shows no change here (Table 1 and Figure 2A and B).

In simulations where the outliers are distributed on both sides the BCT method gives limits closer to the theoretical limits for the normal and square root normal distributions. The FDB method estimates limits close to the theoretical limits for the distributions (χ^2 [df = 8]; χ^2 [df = 1, ncp = 10]) (Figure 2A and B, Table 1 and Supplementary Tables 3 and 4). In the χ^2 (df = 1, ncp = 10) distribution the FDB method detects fewer outliers resulting in an underestimation of the lower limit (Figure 2B and Table 1).

In more skewed distributions such as the log-normal distribution both methods have problems defining outliers. The violin plot (Figure 3A) shows that without outliers the use of the FDB method results in an underestimation

Table 1: Mean and root mean squared error of 2.5th and 97.5th percentiles after outlier elimination (n = 120).

Distribution/percentiles		No outliers				5% outliers			
		BCT		FDB		BCT		FDB	
Outliers on upper side									
Percentiles		2.5	97.5	2.5	97.5	2.5	97.5	2.5	97.5
Normal (8.04; 11.96)	Mean	8.055	11.947	8.005	11.99	8.193	13.025	8.014	13.044
	RMSE	0.163	0.18	0.092	0.136	0.316	1.12	0.11	1.206
Square root normal (64.642; 143.041)	Mean	64.895	142.765	63.973	143.436	67.322	170.726	64.121	165.006
	RMSE	2.54	4.374	1.227	4.072	5.156	29.087	1.669	26.924
Half-normal (0.031; 2.241)	Mean	0.032	2.284	0.031	2.243	0.04	3.507	0.032	2.975
	RMSE	0.007	0.061	0	0.152	0.018	1.255	0.005	0.943
Log-normal (0.141; 7.099)	Mean	0.15	7.273	0.136	5.223	0.17	27.836	0.136	5.5
	RMSE	0.024	1.424	0.004	3.502	0.046	22.752	0.011	3.284
χ^2 (df = 1) (0.001; 5.024)	Mean	0.001	5.291	0.001	4.164	0.002	12.678	0.001	4.416
	RMSE	0	0.29	0	1.702	0.002	7.613	0.001	1.64
χ^2 (df = 4) (0.484; 11.143)	Mean	0.506	11.285	0.471	10.815	0.6	18.645	0.473	12.287
	RMSE	0.085	0.886	0.005	1.642	0.188	7.814	0.046	3.274
χ^2 (df = 8) (2.18; 17.535)	Mean	2.184	17.581	2.094	17.419	2.443	26.491	2.097	20.214
	RMSE	0.216	1.036	0.01	1.365	0.489	9.456	0.11	5.173
χ^2 (df = 1, ncp = 10) (1.446; 26.237)	Mean	1.556	26.055	1.38	25.963	1.955	36.899	1.389	31.681
	RMSE	0.383	1.736	0.01	1.974	0.787	11.924	0.148	9.125
Outliers on both sides									
Normal (8.04; 11.96)	Mean	8.049	11.943	8.003	11.988	7.74	12.218	7.443	12.513
	RMSE	0.164	0.178	0.111	0.116	0.61	0.554	0.818	0.793
Square root normal (64.642; 143.041)	Mean	64.79	142.639	63.867	143.362	59.426	149.656	53.643	153.133
	RMSE	2.531	4.38	0.97	3.775	10.255	14.287	13.554	16.901
χ^2 (df = 8) (2.18; 17.535)	Mean	2.236	17.617	2.132	17.404	1.769	20.046	1.249	18.639
	RMSE	0.248	1.066	0.015	1.499	0.822	4.605	1.09	3.327
χ^2 (df = 1, ncp = 10) (1.446; 26.237)	Mean	1.502	26.296	1.336	26.154	1.077	29.178	0.46	28.508
	RMSE	0.37	1.571	0.014	2.041	0.941	5.695	1.108	5.664

BCT, Box-Cox transformation; FDB, Freedman-Diaconis binning; RMSE, root mean squared error; ncp, non-centrality parameter.

of the upper limit (Table 1). However, in the presence of outliers, the use of the BCT method results in an overestimation and a large RMSE of the upper limits, whereas the FDB method gives a result similar to the result obtained when outliers are not added. For the χ^2 (df = 8) distribution (Figure 3B), the percentages of outliers eliminated when none are added are similar for both methods. When outliers are distributed on both sides, the determination of the upper limit for the FDB method has greater probability mass near the theoretical limit (Table 1). At the lower side, the BCT method eliminates more outliers, which results in a lower limit closer to the theoretical limit.

Adding more extreme outliers using the method described by Horn et al. [18], leads to similar results. However, here the BCT estimation of the lower limits deviates more from the distribution percentiles than with less extreme outliers (Supplementary Figures 1–3 and Supplementary Tables 5 and 6). Adding outliers only on the lower side of skewed distributions shows that the BCT method estimates the lower limits slightly better than the FDB method. As seen for other distributions this then affects the estimation of the other limit, which then deviates more from the distribution percentiles than the FDB method (Supplementary Figures 4 and 5 and Supplementary Tables 7 and 8).

Results of simulation experiments with lower sample sizes show essentially similar results. However, for the very small sample size (n = 30) it seems that the FDB method is better at estimating the limits when no outliers are added (Supplementary Tables 9–16).

Clinical data

A comparison of the FDB method with the BCT method on laboratory data is shown in Table 2. The BCT method detects between 0 and 5% outliers and the FDB method detects between 0 and 8% outliers. Clinically significant differences are found for upper limits of alanine transaminase, creatine kinase and γ-glutamyltransferase. The largest difference is found for γ-glutamyltransferase

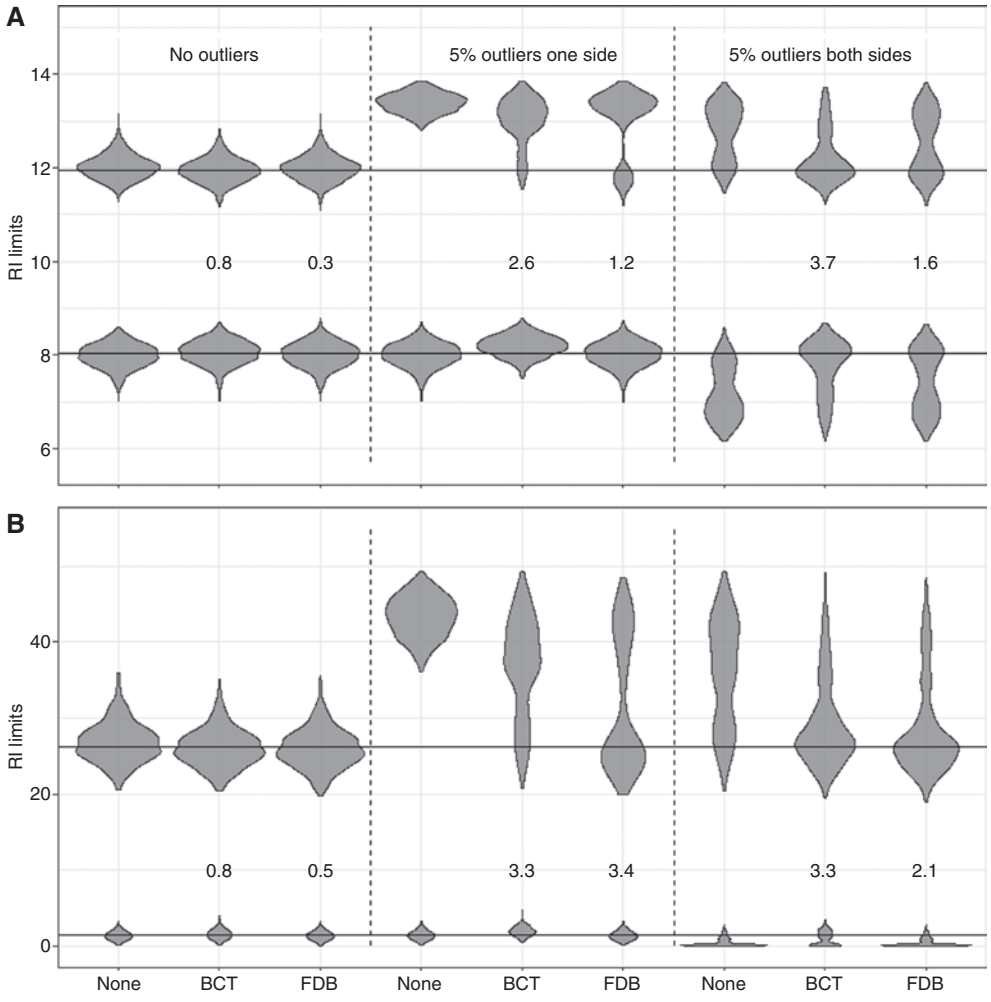

Figure 2: Violin plots of the lower and upper RI limits determined with and without outliers removal.
Each plot shows the result with no added outliers to the left and with 5% outliers added on the upper side in the middle and 5% outliers distributed on both sides to the right. The gray areas (the violin) indicates the distribution and probability density of the determined limits (1000 simulations). The black horizontal lines throughout each panel indicates the theoretical limits (2.5th and 97.5th). The numbers in the middle of the plot indicates how many percent outliers each method on average has detected. (A) Normal distribution, and (B) χ^2 (df = 1, ncp = 10).

(male, >40) (Figure 4A). For this test, the BCT method does not detect any outliers whereas the FDB method detects six. The FDB method evaluates the limits numerically closer to Local and NORIP RIs. The evaluation of the transformation is shown in Supplementary Figure 6. In the case of creatine kinase (male, 18–60) and alanine transaminase the differences between the methods are found at the upper limit, where the FDB method evaluates the limits numerically closer to the Local RIs. For all these RIs (except two γ-glutamyltransferase stratifications) outlier definition according to Dixon's Q test leads to exclusion of fewer outliers than the BCT method (Data not show).

For tests like albumin (Figure 4B), pancreatic amylase, alkaline phosphatase, bilirubin, calcium, creatinine, iron, potassium, sodium and urate the methods does not lead to clinical significant differences in the evaluated limits. However, for some of the skewed distributions like alanine transaminase, pancreatic amylase, alkaline phosphates, creatine kinase, and Iron the outliers removed with the BCT method leads to a small increase in the level of the lower limits.

Repeated detection of outliers on the datasets results in removal of six (one round) and 66 (three rounds) additional outliers with the FDB and BCT methods, respectively. Only γ-glutamyl-transferase (Female, >40) with the FDB method shows a clinically significant change (Supplementary Tables 17 and 18).

General statistics of the datasets are found in (Supplementary Table 19). This table outlines the skewedness of the datasets and the percentiles (2.5 and 97.5) represents the calculated RI if no outliers were defined.

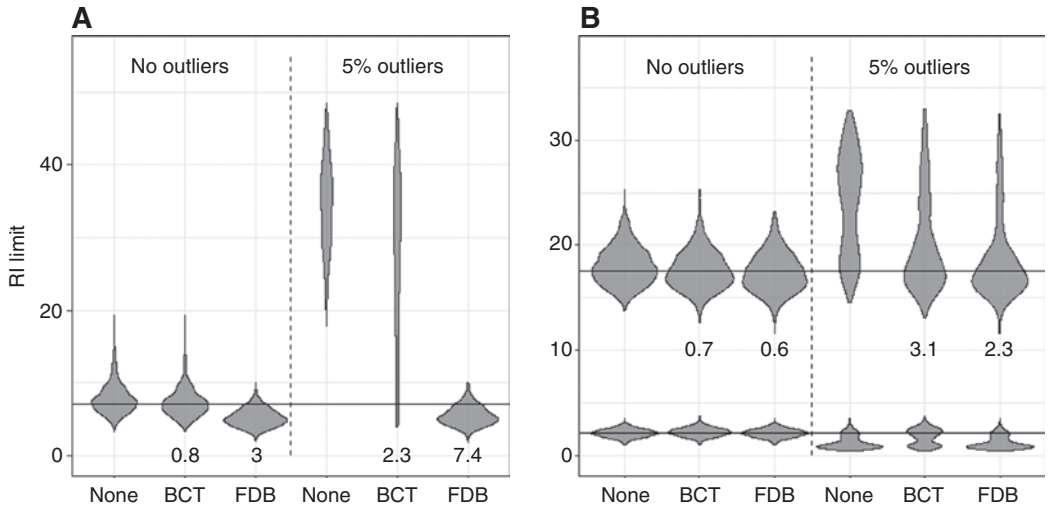

Figure 3: Violin plots of the limits determined with and without outlier removal.
Outline as in Figure 2. (A) Log-normal distribution outliers added above upper limit. Only upper limit shown; and (B) χ^2 (df = 8) distribution outliers added on both sides.

Table 2: Comparison of calculated reference interval limits on adults.

Test	Stratification (gender/age)	Local RI	n total	BCT			FDB		
				Outliers detected	Lower limit	Upper limit	Outliers detected	Lower limit	Upper limit
Alanine transaminase, U/L[a]	Female	10–35	1083	23	9.1	51.9	26	8.3	44.1
	Male	10–50	982	19	12.4	90.8	35	11.4	68.7
Albumin, g/L	18–39	36–48	594	5	33.6	45.6	2	33.6	45.9
	40–70	36–45	788	8	33.9	43.7	6	33.9	43.8
	>70	34–45	162	3	31.3	42.1	3	31.3	42.1
Pancreatic, aamylase, U/L		10–65	730	14	12.2	50.4	4	10.8	50.7
Alkaline phosphatase, U/L		35–105	1524	35	38.5	114.4	13	35.0	115.1
Bilirubin, mmol/L		5–25	962	7	3	22.2	15	3.0	19.9
Calcium, mmol/L[b]		2.20–2.55	1254	21	2.24	2.56	12	2.23	2.57
Creatine kinase, U/L[c]	Female	50–150	235	9	32.91	207.1	7	26.5	182.3
	Male, 18–60	50–270	165	6	59.8	501.2	13	45.6	311.4
	Male, >60	50–200	54	0	27.91	280	0	27.9	280
Creatinine, μmol/L	Female	45–90	1182	29	47.8	88.8	5	46.5	93.7
	Male	60–105	1052	22	62.7	110.2	7	61.4	112.8
Iron, μmol/L		9–34	598	28	6.4	29.7	0	4	33.0
γ-Glutamyl-transferase, U/L	Female, 18–40	10–45	96	0	9.3	49.5	1	9.2	44.6
	Female, >40	10–75	116	0	8.9	107.4	4	8.8	83.0
	Male, 18–40	10–80	96	0	10.9	99.0	3	10.8	62.4
	Male, >40	15–115	105	0	13.4	385.8	6	13.1	123.1
Potassium, mmol/L[d]		3.5–4.6	2125	42	3.41	4.61	3	3.35	4.68
Sodium, mmol/L[e]		137–145	2102	51	137	144.4	13	136.3	144.6
Urate, μmol/L	Female, 18–50	0.16–0.35	119	2	0.14	0.38	0	0.14	0.39
	Female, >50	0.16–0.40	89	3	0.17	0.49	2	0.15	0.44
	Male	0.23–0.48	197	1	0.22	0.55	1	0.22	0.55

Differences between local and NORIP reference intervals. [a]Female, 10–45 U/L; male, 10–70 U/L. [b]2.15–2.51 mmol/L. [c]Female, 35–210 U/L; male, age 18–49, 50–400 U/L; male, age ≥50, 40–280 U/L. [d]3.5–4.4 mmol/L. [e]137–144 mmol/L. Outliers are defined by the BCT or FDB methods. Local reference intervals (RI) are given in column 2. These RIs are based on NORIP RIs [8]. BCT, Box-Cox transformation; FDB, Freedman-Diaconis binning.

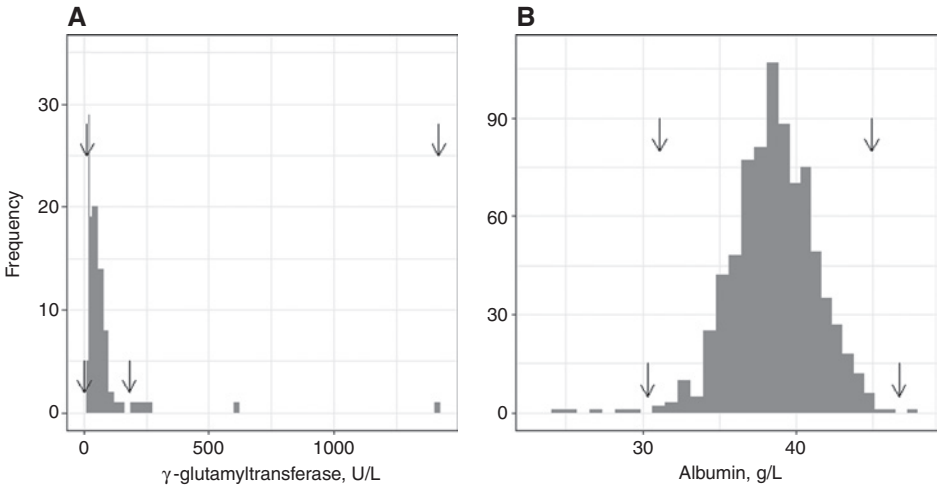

Figure 4: Histograms of laboratory data containing outliers.
The arrows in the top of the histograms indicates outlier fences as defined with the BCT method. The arrows in the bottom of the histogram indicate the first empty bin smaller or larger than the median, defining the proposed transformation independent outlier fences. (A) Histogram of γ-glutamyltransferase results (n = 105); and (B) histogram of albumin results (n = 788).

Discussion

We have presented a simple transformation independent method for outlier definition. The method uses the concept of drawing a histogram of the data and defines the outlier fences as the first bins above and below the median not containing any data. The bin sizes are defined using a modified Freedman-Diaconis rule. The modification allows for defining individual bin sizes above and below the median. This method is simple, reproducible, and intuitively understandable; it gives results resembling what a clinical biochemistry professional would reach after inspecting a histogram. Evaluation of simulation experiments comparing the FDB method to the BCT method show that the FDB method, in most cases, is better at defining outliers when added on one side. In this case reference limits are more precise both at the upper and lower limit. When outliers are added on both sides, the BCT method is better at the lower side, whereas the methods are approximately equally good at the upper side. However, the results indicate that the BCT elimination of outliers at the lower side is a result of addition of outliers on the upper side. In simulation experiments without outliers added, the FDB method seems to define a few more outliers resulting in a smaller deviation from the theoretical limits. Using smaller samples similar results are found, only in very small samples the FDB method seems to be better at including true points when outliers are not added.

The results from the evaluation of laboratory data confirms several of the results from the simulation experiments. When clinically significant differences are found, the limits determined with the FDB method are numerically closer to the NORIP limits. Similarly, the BCT method influences the level of the lower limit when used on skewed distribution. Though an indirect datasets can be influence by selection bias it is found useful, as it is a dataset of actual measured values. The difficulties, which the FDB method has in detecting outliers on the lower side, does not influence the clinical evaluation. At least not to the same extent as the BCT method influence on the lower limit.

The apparent advantage of the FDB method may be explained by the fact that the IQR is essentially insensitive to extremes, whereas a transformation of data is influence by outliers. This is especially seen in skewed distributions (like Figure 3A and general statistics of the datasets). The evaluation of the transformation for this distribution (Supplementary Figure 6) shows that the results in this case is not due to bad transformation or remaining kurtosis. The remaining kurtosis has in other simulation experiments been described as the reason for suboptimal results [1]. Furthermore, the FDB method is dependent on the sample size, as the bin size is dependent on sample size. This and the evaluation of simulation experiments with low sample number and of the laboratory data shows that the FDB method seems to perform equally well at different sample sizes.

The FDB method has the statistical limitation that it is not applicable when the underlying distribution is uniform [16]. However, this is not a distribution often found in clinical biochemistry. Additionally, multimodal

densities are problematic for the BCT method as well as for the FDB method [21]. However, as the method are to be used in the tails of distributions this will in most cases be a minor problem. At very large sample sizes, the bin size can become smaller than the test resolution resulting in a fragmented histogram. The solution here is to use the test resolution as the bin size. Finally, the proposed method was only tested on indirect datasets, which do not represent a true random sample. The selection relies on the requisition pattern as an acceptable surrogate for good health status [20].

In conclusion, we find our simple transformation independent outlier detection method as good as or better than the currently recommended methods. It allows for a complete non-parametric and transformation independent calculation of RIs. The method is easy to perform manually in Microsoft Excel®. It can be done by building a frequency table from the median using the bin sizes described in formulas (3) and (4). The first empty bin on each side of the median defines the outlier fences. We therefore propose to use this new method once to get a transformation independent outlier definition.

Author contributions: PAC conceived the idea and study. PAC performed calculation and simulations. MBJ and PAC refined the idea and simulations. PAC reviewed the literature and wrote the first draft. Both authors contributed to subsequent drafts and approval of the final version. All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: None declared.
Employment or leadership: None declared.
Honorarium: None declared.
Competing interests: The funding organization(s) played no role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

References

1. Solberg HE, Lahti A. Detection of outliers in reference distributions: performance of Horn's algorithm. Clin Chem 2005;51:2326–32.
2. Dixon WJ. Analysis of extreme values. Ann Math Stat 1950;21:488–506.
3. Grubbs FE. Sample criteria for testing outlying observations. Ann Math Stat 1950;21:27–58.
4. Stromme JH, Rustad P, Steensland H, Theodorsen L, Urdal P. Reference intervals for eight enzymes in blood of adult females and males measured in accordance with the International Federation of Clinical Chemistry reference system at 37 degrees C: part of the Nordic Reference Interval Project. Scand J Clin Lab Invest 2004;64:371–84.
5. Tozzoli R, Giavarina D, Villalta D, Soffiati G, Bizzaro N. Definition of reference limits for autoantibodies to thyroid peroxidase and thyroglobulin in a large population of outpatients using an indirect method based on current data. Arch Pathol Lab Med 2008;132:1924–8.
6. Erasmus RT, Ray U, Nathaniel K, Dowse G. Reference ranges for serum creatinine and urea in elderly coastal Melanesians. P N G Med J 1997;40:89–91.
7. Eskelinen S, Suominen P, Vahlberg T, Lopponen M, Isoaho R, Kivela SL, et al. The effect of thyroid antibody positivity on reference intervals for thyroid stimulating hormone (TSH) and free thyroxine (FT4) in an aged population. Clin Chem Lab Med 2005;43:1380–5.
8. Rustad P, Felding P, Franzson L, Kairisto V, Lahti A, Martensson A, et al. The Nordic Reference Interval Project 2000: recommended reference intervals for 25 common biochemical properties. Scand J Clin Lab Invest 2004;64:271–84.
9. Tukey JW. Exploratory data analysis. Reading, MA: Addison-Wesley, 1977:688.
10. Bjerner J, Theodorsson E, Hovig E, Kallner A. Non-parametric estimation of reference intervals in small non-Gaussian sample sets. Accred Qual Assur 2009;14:185–92.
11. Hoaglin DC, Iglewicz B, Tukey JW. Performance of some resistant rules for outlier labeling. J Am Stat Assoc 1986;81:991–9.
12. Solberg HE. The theory of reference values Part 5. Statistical treatment of collected reference values. Determination of reference limits. J Clin Chem Clin Biochem 1983;21:749–60.
13. Box GE, Cox DR. An analysis of transformations. J R Stat Soc Series B (Methodological) 1964;26:211–52.
14. Harris EK, Boyd JC. Statistical bases of reference values in laboratory medicine. New York: M. Dekker, 1995:xiv, 361.
15. Horn PS, Pesce AJ, Copeland BE. A robust approach to reference interval estimation and evaluation. Clin Chem 1998;44:622–31.
16. Freedman D, Diaconis P. On the histogram as a density estimator – L2 theory. Z Wahrscheinlichkeit 1981;57:453–76.
17. CLSI. Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline – third edition. CLSI document EP28 – A3c ed. Wayne, PA, USA: CLSI (Clinical Laboratory Standards Institute), 2010.
18. Horn PS, Feng L, Li Y, Pesce AJ. Effect of outliers and nonhealthy individuals on reference interval estimation. Clin Chem 2001;47:2137–45.
19. Patterson N. A robust, non-parametric method to identify outliers and improve final yield and quality. CS MANTECH Conference; April 23rd–26th, 2012; Boston, MA, USA, 2012.
20. Lykkeboe S, Nielsen CG, Christensen PA. Indirect method for validating transference of reference intervals. Clin Chem Lab Med 2018;56:463–70.
21. Knuth KH. Optimal data-based binning for histograms, 2006. arXiv:physics/0605197 [physicsdata-an].

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/cclm-2018-0025).